Mathematics 376 – Probability and Statistics II
Solutions for Midterm Exam 1 – March 11, 2010

I. In a random poll of female students at BigState University, 390 out of 500 sorority women were in favor of lengthening the spring break, while 64 out of 100 non-sorority women were in favor.

A) (15) Give a 99% confidence interval for the difference of the proportions favoring lengthening spring break.

Solution – The point estimators for the two population proportions are $\frac{390}{500} = .78$ and $\frac{64}{100} = .64$. By the large sample formulas, $z_{.005} \doteq 2.575$, and the confidence interval is

$$(.78 - .64) \pm 2.575\sqrt{\frac{(.78)(.22)}{500} + \frac{(.64)(.36)}{100}} \doteq .14 \pm .1325.$$

B) (5) Now suppose a new poll of equal numbers of women from both groups is being planned. How many women should be polled so that error of estimation is no larger than .08 according to a two standard error bound? (Be "conservative.")

Solution – We want to choose the number n of women to be polled so that

$$2\sqrt{\frac{(.5)(.5)}{n} + \frac{(.5)(.5)}{n}} \leq .08.$$

This says $n \geq 312.5$, so $n \geq 313$.

II. Two methods for teaching reading were applied to two randomly selected groups of school children, and compared by means of a common test. The sample means ($\overline{y}$) and variances (s^2) of the test scores for the two groups are as follows:

Statistic	Method 1	Method 2
Number of Children	11	14
$\overline{y}$	64	69
s^2	20	15

A) (10) Find a 95% confidence interval for σ_1^2 (the variance of the test scores for students taught with Method 1. What assumptions are you making?

Solution – We need to assume that the test scores are from a normally distributed population. Then using the χ^2 table, the confidence interval is

$$\frac{(n_1 - 1)S_1^2}{\chi^2_{.025}} = \frac{10 \cdot 20}{20.4831} \leq \sigma_1^2 \leq \frac{10 \cdot 20}{3.24697} = \frac{(n_1 - 1)S_1^2}{\chi^2_{.975}}$$

or

$$9.764 \leq \sigma_1^2 \leq 61.596.$$

B) (20) Find a 95% confidence interval for $\mu_1 - \mu_2$ (the difference in the population mean test scores with the two methods). What assumptions are you making?

Solution – We need to assume the samples are independent from normal populations with equal variance. (Note that the sample variance from the students taught with Method 2 is in the interval from part A, which is a limited indication that the equality of variances is reasonable.) From the small sample formulas, we have that the pooled estimator for the variance is

$$S_p^2 = \frac{10 \cdot 20 + 13 \cdot 15}{23} \doteq 17.1739.$$

Then since $t_{.025}(23) = 2.069$,

$$\mu_1 - \mu_2 = (64 - 69) \pm 2.069 \cdot \sqrt{17.1739} \cdot \sqrt{1/11 + 1/14} \doteq -5 \pm 3.4547.$$

C) (5) Do your results support the conclusion that Method 2 is a superior method for teaching reading? Explain.

Solution – Yes, since the interval in part B consists of only negative numbers, there is evidence (at the 95% confidence level) to say that Method 2 yields superior results.

III. (15) A random sample of size 100 was taken from a normal population with $\sigma^2 = 8$. A confidence interval for the population mean was given as $6 \pm .047$. What was the confidence coefficient associated with this interval (that is the α value in the $(1 - \alpha) \times 100\%$)?

Solution – Since the number of samples was $100 > 30$, the large sample formulas would be used:

$$.047 = z_{\alpha/2} \frac{\sqrt{8}}{\sqrt{100}},$$

so $z_{\alpha/2} = .47/\sqrt{8} \doteq .166$. Using the closest z value from the standard normal table, this says $\alpha/2 \doteq .4325$. Hence $\alpha \doteq .865$. This was a 13.5% confidence interval. (Not one of the standard values! And actually, this is a somewhat nonsensical value because this problem contained an unfortunate typo – I had meant to make the reported interval 6 ± 0.47 but I switched the 0 and the decimal point and I did not catch this(!))

IV. from a population described by the pdf:

$$f(y) = \begin{cases} \frac{2y}{\theta^2} & \text{if } 0 \le y \le \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y_{(n)}$ be the sample maximum and let $U = \frac{1}{\theta} Y_{(n)}$.

A) (10) Find a constant c so that $\widehat{\theta} = cY_{(n)}$ is an unbiased estimator for θ.

Solution – We have

$$F(y) = \begin{cases} 0 & \text{if } y < 0 \\ \frac{y^2}{\theta^2} & \text{if } 0 \le y \le \theta \\ 1 & \text{if } y > \theta. \end{cases}$$

Therefore from the formulas for order statistics, the PDF for the sample maximum is

$$f_{(n)}(y) = n \left(\frac{y^2}{\theta^2}\right)^{n-1} \frac{2y}{\theta^2}.$$

This implies

$$\begin{aligned} E(Y_{(n)}) &= \int_0^\theta y \cdot n \left(\frac{y^2}{\theta^2}\right)^{n-1} \frac{2y}{\theta^2}\, dy \\ &= \int_0^\theta 2n \frac{y^{2n}}{\theta^{2n}}\, dy \\ &= \frac{2n}{(2n+1)} \frac{y^{2n+1}}{\theta^{2n}}\bigg|_0^\theta \\ &= \frac{2n\theta}{2n+1}. \end{aligned}$$

This shows that $Y_{(n)}$ is a biased estimator for θ, but $\widehat{\theta} = \frac{2n+1}{2n} Y_{(n)}$ is unbiased.

B) (10) Show that U has CDF given by

$$F_U(u) = \begin{cases} 0 & \text{if } u < 0 \\ u^{2n} & \text{if } 0 \le u \le 1 \\ 1 & \text{if } u > 1 \end{cases}$$

Solution – By definition the CDF of U is

$$\begin{aligned} F_U(u) &= P(U \le u) \\ &= P(Y_{(n)} \le \theta u) \\ &= \begin{cases} 0 & \text{if } u < 0 \\ \int_0^{\theta u} n \left(\frac{y^2}{\theta^2}\right)^{n-1} \frac{2y}{\theta^2}\, dy & \text{if } 0 \le u \le 1 \\ 1 & \text{if } u > 1. \end{cases} \\ &= \begin{cases} 0 & \text{if } u < 0 \\ u^{2n} & \text{if } 0 \le u \le 1 \\ 1 & \text{if } u > 1 \end{cases} \end{aligned}$$

as we wanted to show.

C) (10) Note that the distribution of U does not depend on θ. Use this information to derive a 95% *upper* confidence bound on θ from the samples.

Solution – $P(U \geq u) = 1 - P(U \leq u) = .95$ when $P(U \leq u) = .05$. By part B, this says $u = (.05)^{1/(2n)}$. Hence

$$P\left(\frac{1}{\theta}Y_{(n)} \geq (.05)^{1/(2n)}\right) = .95,$$

which is equivalent to

$$P\left(\theta \leq \frac{Y_{(n)}}{(.05)^{1/(2n)}}\right) = .95.$$

So the 95% upper confidence bound is

$$\widehat{\theta_U} = \frac{Y_{(n)}}{(.05)^{1/(2n)}}.$$

Extra Credit.

A) (5) (Multiple Choice, and explain.) A survey was conducted to determine what features users want included in cell phone service. The results said a 95% confidence interval showed that $73\% \pm 4\%$ of users wanted email services. What is meant by the $\pm 4\%$?

1) As many as 4% of those polled may change their minds after the time the survey was conducted.
2) The probability is .95 that the true population percentage favoring email service is in the interval $(69\%, 77\%)$.
3) It would be unlikely to get the observed sample proportion if the true population percentage was outside the interval $(69, 77)$.
4) Only 4% of cellphone users were polled.

Solution – Only 3 is correct. 1 and 4 are clearly nonsense. 2 is the incorrect interpretation of confidence intervals that we discussed in Lab 1 (recall that there is no chance involved in the location of the population percentage relative to the interval – the chance is in the sampling process that produces the confidence interval).

B) (5) Let $Z_1, \ldots, Z_5$ be independent samples from a $N(0,1)$ population, and define:

$$W = \frac{2Z_1}{\sqrt{Z_2^2 + \cdots + Z_5^2}}.$$

What is the distribution of W^2? Explain.

Solution – By definition, W has a t-distribution with 4 degrees of freedom. Therefore

$$W^2 = \frac{(Z_1^2)/1}{(Z_2^2 + \cdots + Z_5^2)/4}$$

is a ratio of χ^2's divided by their numbers of degrees of freedom. Thus W has an F-distribution with 1 numerator degree of freedom and 4 denominator degrees of freedom.